

## LA-UR-21-31693

Approved for public release; distribution is unlimited.

Title: Implementing Artificial Intelligence Technology at a Major Library

Author(s): Ali, Alee Rizwan

Intended for: presentation

Issued: 2021-11-30

---

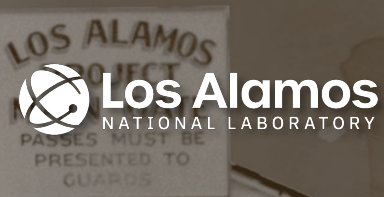
**Disclaimer:**

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

LOS ALAMOS  
PROJECT  
MAIN GATE  
PASSES MUST BE  
PRESENTED TO  
GUARDS

# NATIONAL SECURITY RESEARCH CENTER





# Implementing Artificial Intelligence Technology at a Major Library



Rizwan Ali  
Director, NSRC  
Los Alamos National Laboratory  
[rizwan@lanl.gov](mailto:rizwan@lanl.gov)



Managed by Triad National Security, LLC for the U.S. Department of Energy's NNSA



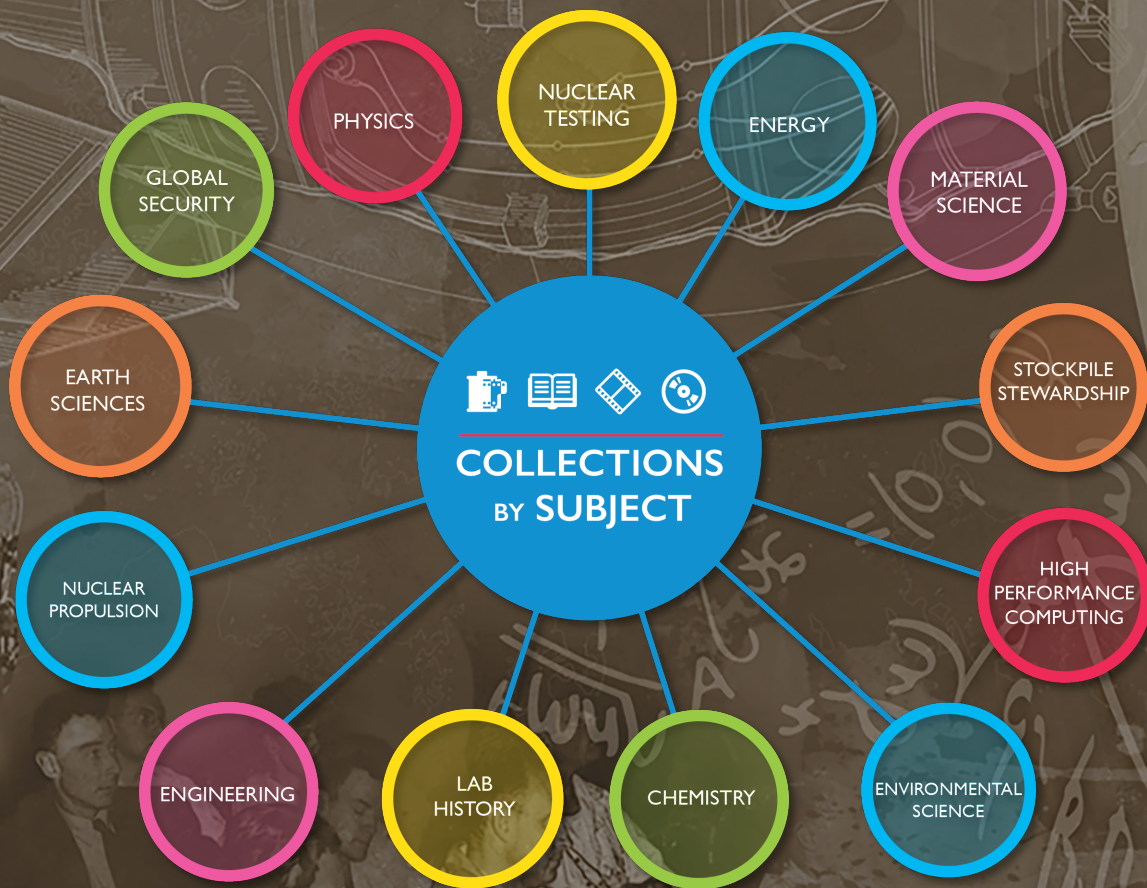
# About Us



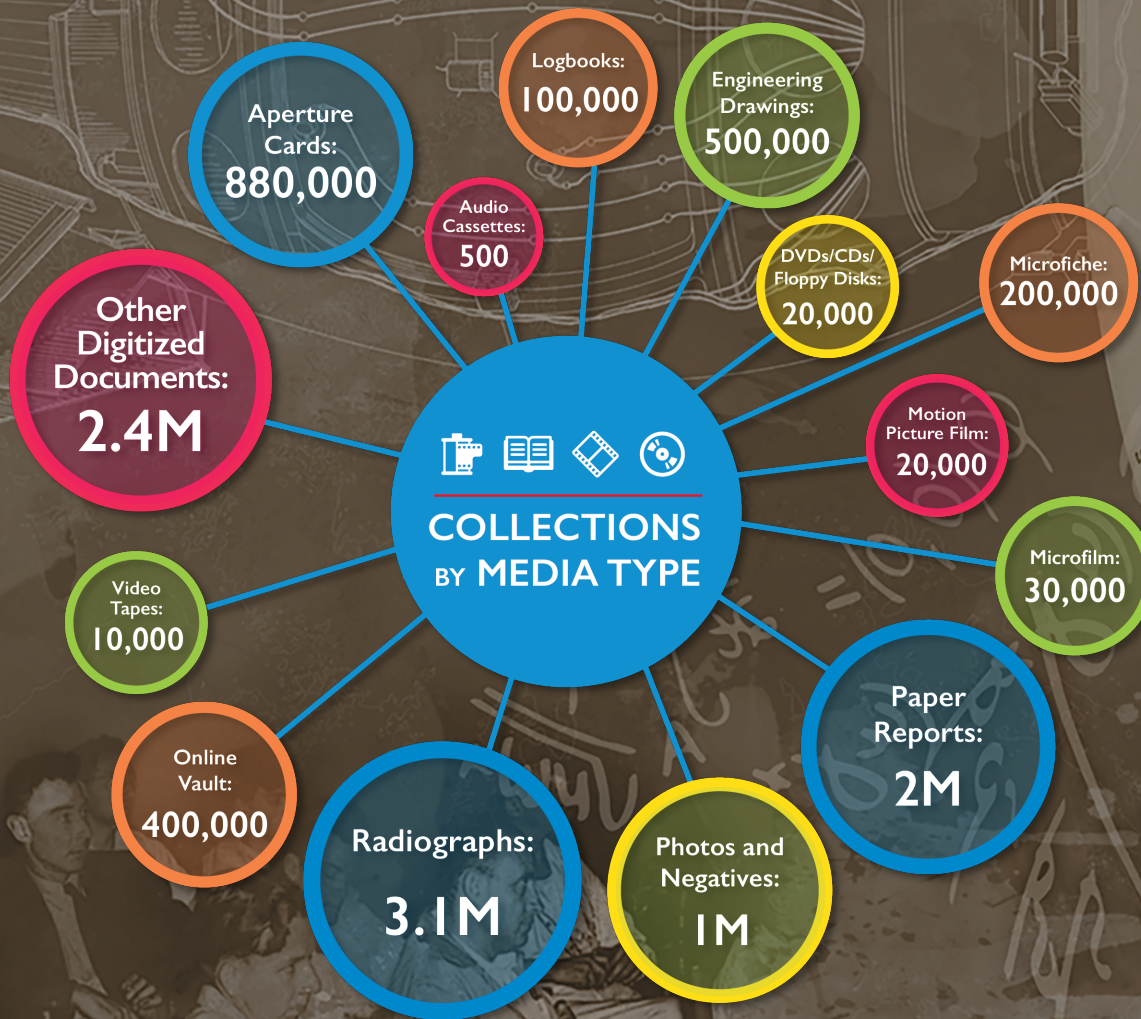
- NSRC is Los Alamos National Laboratory's classified library.
- Lineage dates to the **Technical Library** formed by **J. Robert Oppenheimer** in 1943 as part of the Manhattan Project.
- Library is the end result of **decades of consolidation** of mini libraries and mini archives at LANL.
- Houses **75+ years** of scientific and **engineering** research, designs, procedures, videos, photos, and other reports.



*The NSRC offers services similar to major university research libraries*







# Digitizing Efforts



- Less than 10% of the physical collection is digitized.
- Decades-long project to digitize the following media types:

→ Paper

→ Audio

→ Engineering drawings

→ Notebooks

→ Microfilm

→ Negatives

→ Microfiche

→ Card catalog

→ Photos

→ Aperture cards

→ Video

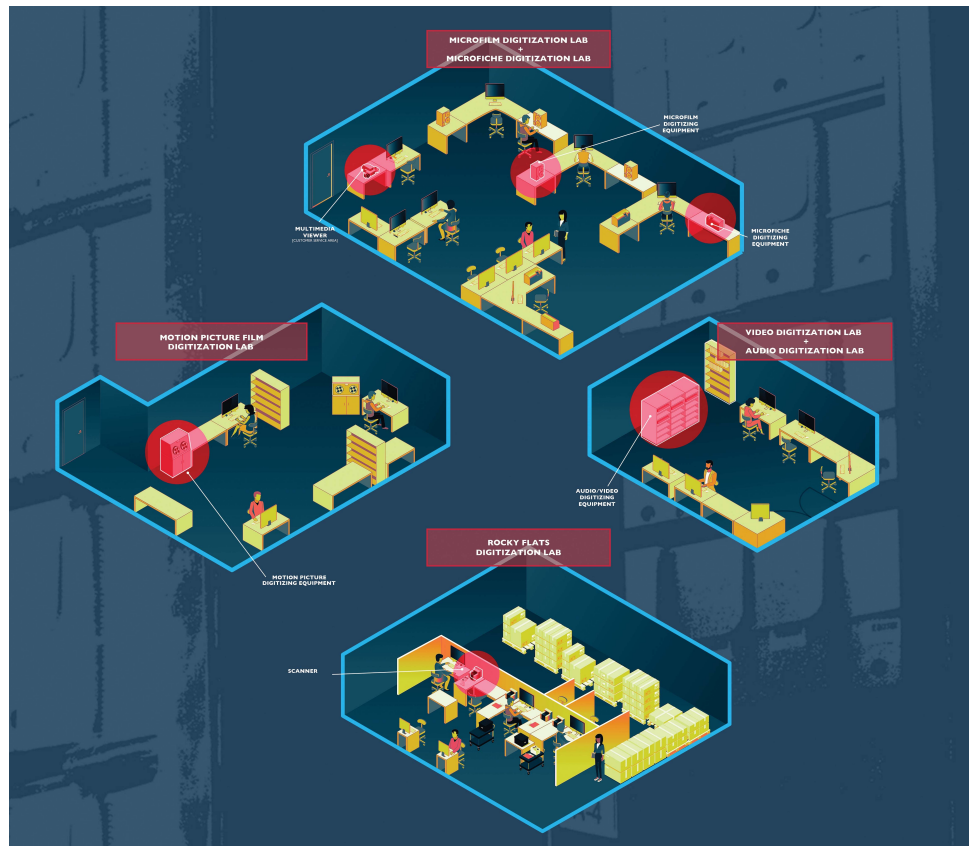
→ Motion picture film



*NSRC has established high-speed digitizing labs for several media types*



# New high-speed digitizing labs



- Seven new digitization labs in 2020 & 2021:

- Video Tape Digitization Lab
- Audio Tape Digitization Lab
- Microfilm Digitization Lab
- Microfiche Digitization Lab
- Motion Picture Film Digitization Lab
- Rocky Flats Digitization Lab
- Specialized Digitization Lab

- Two new labs planned for 2022

- Paper Digitization Lab
- Indexing Lab

*Rapidly expanding high-speed digitizing labs presents indexing/cataloging challenge*



# Cataloging / Indexing Efforts

- Digital Collections

- Wide variety of metadata and indices ranging from nonexistent to rudimentary to robust
- *Search functions range from nonexistent to difficult-to-use*

- Physical Collections

- *Poor to nonexistent metadata*
- *Limited indices*

- Cataloging rate:

- *Currently using **fully manual** process*
- *Process takes 10-30 minutes per document*
- *Total number of **digitized files** growing rapidly*
- *For just one of our digital collections, process will take 400+ years*

Cataloging snapshot of just one digitized collection	
Quantity	2.4 million
FTEs	1.5
Rate per month	486
Years to complete	412

*The backlog in cataloging/indexing cannot be met with manual approach*

# Automated cataloging using Artificial Intelligence / Machine Learning (Titan on the Red)



- **Basic requirements:**
  - Automatically extract metadata from digitized documents
  - Perform natural-language search
  - Ingest data from various repositories (SharePoint, shared drives, etc.).
- **Unique requirements for classified library holdings:**
  - Enforce security classification rules
  - Enforce 'need to know' protocols
  - Utilize public domain, commercial, and Los Alamos ontologies
- **Finding the companies:**
  - Not easy to find...contacted over a dozen companies
  - Unique problem set...but not that unique
  - US Intelligence Community has very similar requirements and several companies that specialize in addressing IC's needs
  - Companies found through personal contacts in the IC
- **Phases:**
  - Phase 1: 6-month pilot/demo unclassified data
  - Phase 2: Currently being implemented on the Los Alamos classified networks




## Goals

- Provide an integrated search across document repositories
- Improve time to discovery
- Reduce data uncertainty
- Ensure secure access

## Capabilities

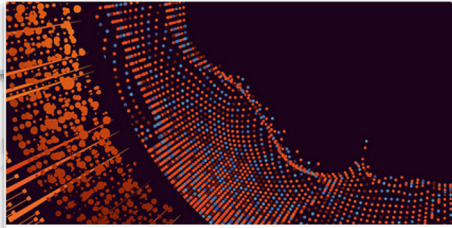
- Full text ingestion and search
- Auto-metadata extraction
- Natural Language Understanding/Processing



**Data Ingest Tool**

AI-Enabled ETL for Ultra Large Scale Digitization

[Launch](#)



**Digital Archives**

Precision Search & Recall for Digitized Assets

[Launch](#)





## Ingestion and Metadata Extraction

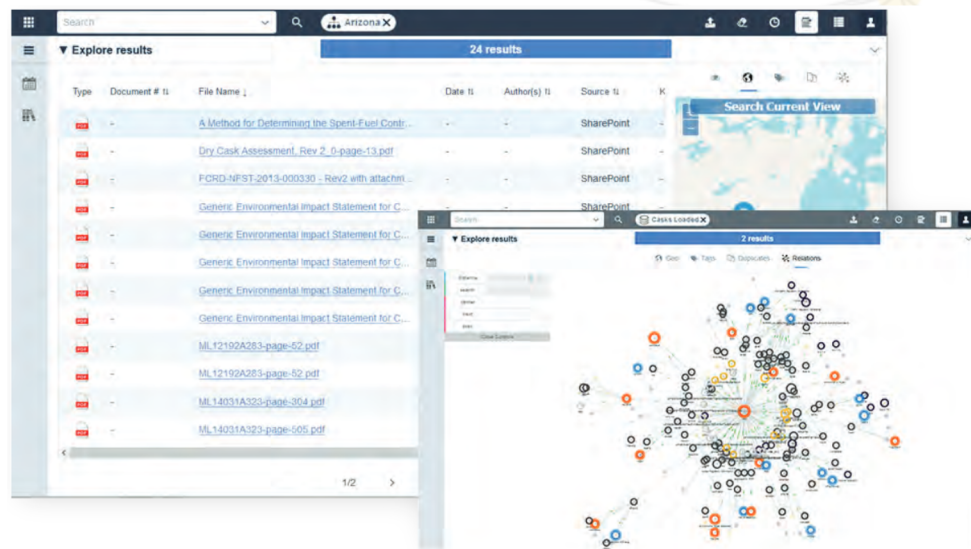
- AI-enabled Extract, Transform, Load (ETL)
- Machine learning based Optical Character Recognition (OCR)
  - Tesseract to recognize pre-1984 fonts
- Machine learning to recognize metadata
- Barcodes, ID #, authors, abstracts, TOC, etc.

# Natural Language Processing



## Natural Language Processing

- Auto-generation of metadata
- Categorizes documents
- Reveals semantic relationships
- Enables discovery of hidden / inferred substantive content



tt\_cdp\_01072020



## Roadmap

**2016**

### Technology Evals.

*Evaluation of 7 tools*

**2017**

### Pilot with Palantir

*-not selected-*

*Test for functionality and classified implementation*

**2020**

### Pilot with Titan

*Test for functionality and classified implementation*

**2021-2022**

### Classified Environment

*Implement classified license and infrastructure*

**2023**

### Expansion

*Connect to additional data sources*

**2024+**

### Expansion and Analysis

*Evaluate analytic capabilities to leverage relationships among data*

- National Security Research Center,  
Director: Rizwan Ali, [rizwan@lanl.gov](mailto:rizwan@lanl.gov)
- National Security Research Center,  
Titan on the Red (AI/ML) Project  
Manager: Julie Maze, [jmaze@lanl.gov](mailto:jmaze@lanl.gov)



LOS ALAMOS  
PROJECT  
MAIN GATE  
PASSES MUST BE  
PRESENTED TO  
GUARDS

# NATIONAL SECURITY RESEARCH CENTER

